Chapter 2

The cichlid genome project

2.1 Introduction

2.1.1 Publication note

The work described in section 2.1.2 and sections 2.3.1 to 2.3.3 was previously published in Brawand, Wagner, Li, Malinsky *et al.*, 2014 [93]. This publication was the result of over five years of work by the 'Cichlid Genome Consortium' and its 76 members. As always, results obtained by others are indicated in the text; everything else is my own work.

2.1.2 Five reference genomes

Recognising the potential of the East African cichlid radiations to generate many important insights into the genetic basis of speciation and functional diversification, a Cichlid Genome Consortium led by the Broad Institute generated draft reference genome assemblies for five species. The genomes were selected to provide one reference for each of five major lineages of East African cichlids, focussing on Lakes Tanganyika, Malawi, and Victoria (Figure 2.1) The reference genomes, listed in Table 2.1, are used throughout the rest of this thesis. *Oreochromis niloticus*, commonly known as Nile tilapia, is a riverine cichlid that shared a common ancestor with the highly radiating lake cichlids approximately 25-50 million years ago and so provides an outgroup for the study of their evolution. *Neolamprologus brichardi*, a reef-dwelling planktivore species, is representative of Lamprologini, the most numerous tribe of Lake Tanganyika comprising 79 endemic species. *Astatotilapia burtoni* is a Tanganyikan representative of the tribe Haplochromini. It is one of the few species able to cross the lake-river boundaries and therefore is also found in the rivers throughout the Lake Tanganyika catchment. *Metriaclima zebra* is a representative of the rock-dwelling 'mbuna' lineage of Lake Malawi, an exemplar of a rapidly speciating genus, and is a highly specialised lake species, an algae scraper. Finally, *Pundamilia nyererei*, a widely distributed reef-dwelling planktivore, is a representative of the young cichlid radiation of Lake Victoria.

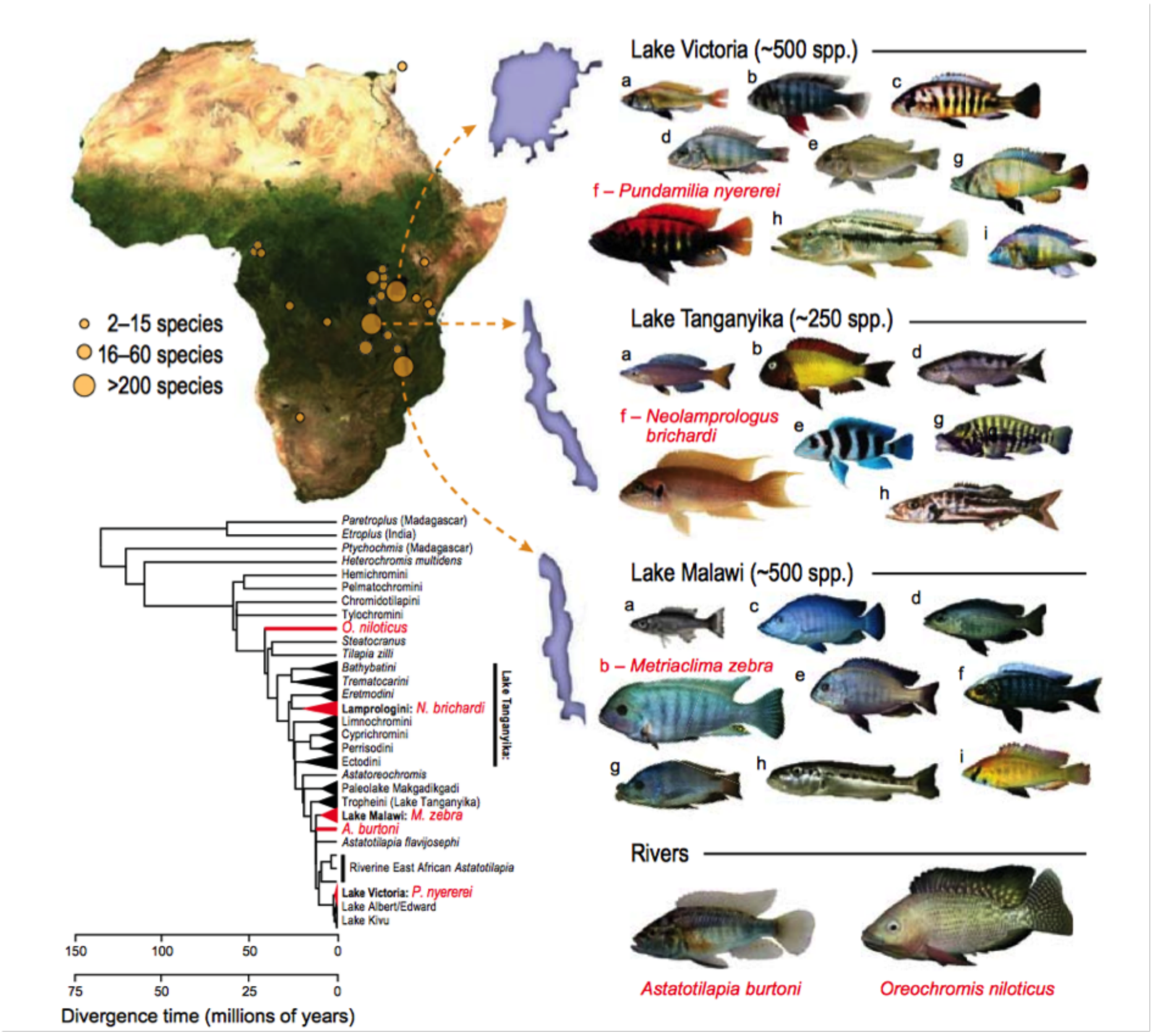


Fig. 2.1 **East African radiation of cichlid fish. Top left:** A map showing the location of African lakes in which cichlids have radiated. **Bottom left:** A phylogenetic tree showing the relationship between the sequenced species (red) and other cichlid lineages, with timescales reflecting two different estimates. **Right:** The sequenced species (red) and examples of major ecotypes in each lake: a, pelagic zooplanktivore; b, rock-dwelling algae scraper; c, paedophage (absent from Lake Tanganyika); d, scale eater; e, snail crusher; f, reef-dwelling planktivore; g, lobe-lipped insect eater; h, pelagic piscivore; i, ancestral river-dweller also found in lakes. Figure from [93].

Table 2.1 **Versions of cichlid genome assemblies used in this thesis**.

Species	Broad Institute assembly	URL used to download
M. zebra	MetZeb1.1_prescreen	http://www.broadinstitute.org/ftp/pub/assemblies/fish/M_zebra/MetZeb1.1_prescreen/M_zebra_v0.assembly.fasta
P. nyererei	PunNye1.0	http://www.broadinstitute.org/ftp/pub/assemblies/fish/P_nyererei/PunNye1.0/P_nyererei_v1.assembly.fasta.gz
A. burtoni	HapBur1.0	http://www.broadinstitute.org/ftp/pub/assemblies/fish/H_burtoni/HapBur1.0/H_burtoni_v1.assembly.fasta.gz
N. brichardi	NeoBri1.0	http://www.broadinstitute.org/ftp/pub/assemblies/fish/N_brichardi/NeoBri1.0/N_brichardi_v1.assembly.fasta.gz
O. niloticus	Orenil1.1	http://www.broadinstitute.org/ftp/pub/assemblies/fish/tilapia/Orenil1.1/20120125_MapAssembly.anchored.assembly.fasta

In addition to generating the reference genomes, the project team obtained RNA sequence data from multiple tissues in each species and used it to generate high quality annotation of protein coding genes. Comparisons between the annotated cichlid genomes and also utilising the existing genomes of other teleost fish (medaka, stickleback, tetraodon, and zebrafish) then provided interesting insights into the genomic changes underlying evolution of cichlids in East Africa. Details of these findings are described in the Cichlid genome consortium publication [93]. Here I briefly describe three highlights. First, the consortium members discovered 4.5- to 6-fold increase in the rate of gene duplications in the ancestors of the rapidly radiating lake cichlids and of the haplochromines. Duplicate genes can evolve new functions through neo- or subfunctionalisation [3] and thus facilitate adaptation and speciation. Second, the rate of evolution of genes associated with changes in jaw morphology is accelerated in haplochromines, providing evidence of repeated positive selection on these genes. And third, 625 noncoding regions were found to have a significantly accelerated rate of substitution in one or more of the East African lake cichlids, but strong sequence conservation across teleosts. Laboratory experiments indicated the ability of these non-coding sequences to alter the strength and patterns of expression of nearby protein coding genes, further strengthening the evidence for their function in cichlid evolution.

Another type of gene regulation, regulation by microRNAs (miRNAs), could also play an important role in cichlid evolution. I worked under the supervision of Eric Miska at the University of Cambridge to generate a miRNA annotation for each of the five reference genomes, and predictions of genes whose expression may be regulated by the miRNAs (target genes). I also categorised miRNAs in terms of sequence novelty, and assessed the hypothesis that accelerated rate of change in pairing between miRNAs and target genes could have contributed to the divergence of the five cichlid species sequenced by the consortium.

2.2 Background

2.2.1 The nature and functions of microRNAs

Like protein coding genes, animal microRNAs are transcribed by RNA polymerase II. After transcription, a ∼70 to ∼100bp long section of the primary transcript (pri-miRNA) must fold into a so-called 'hairpin' structure (Figure 2.2), the hallmark of microRNAs. The hairpin is then excised from the flanking sequence and exported out of the nucleus. This so-called hairpin precursor (pre-miRNA) is further processed in the cytoplasm. First, its terminal loop is cleaved off, resulting in a ∼22bp double-stranded RNA molecule (Figure 2.2). The two strands separate and one associates with the microRNA-induced silencing complex (miRISC). The other strand is degraded [21, 94]. The selection of the strand to be loaded into miRISC is based on the thermodynamics of the double-stranded RNA molecule. The strand more often detected with miRISC is called mature miRNA, and the more often degraded strand in called miRNA* [94].

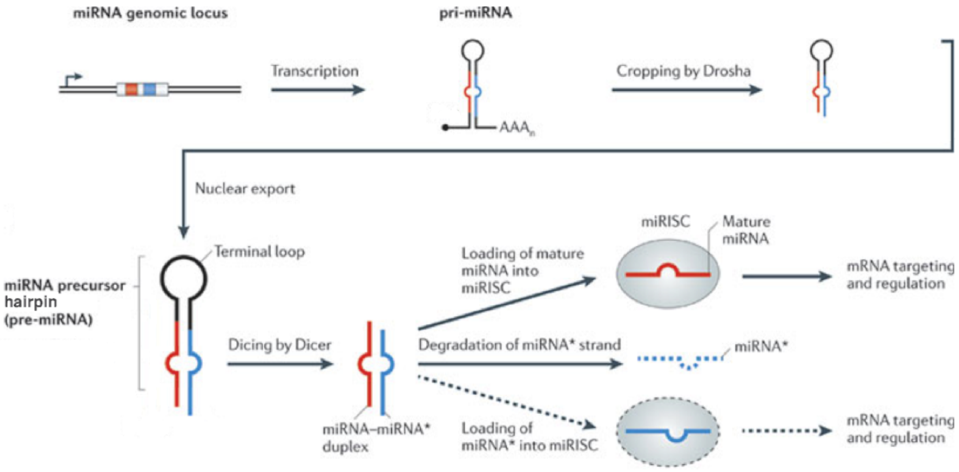

miRNA genomic locus
pri-miRNA
Transcription
Cropping by Drosha
AAA$_n$
Nuclear export
Terminal loop
miRNA precursor hairpin (pre-miRNA)
Dicing by Dicer
miRNA–miRNA* duplex
Loading of mature miRNA into miRISC
miRISC
Mature miRNA
mRNA targeting and regulation
Degradation of miRNA* strand
miRNA*
Loading of miRNA* into miRISC
mRNA targeting and regulation

Fig. 2.2 **Biogenesis of microRNAs.** After transcription, a part of the RNA molecule folds into a hairpin structure. Still in the nucleus, regions flanking the hairpin are cut off by the RNAase type III enzyme Drosha. After being exported out of the nucleus, the miRNA precursor hairpin (pre-miRNA) itself is cleaved in another reaction catalysed by the RNAase type III enzyme Dicer. The resulting miRNA-miRNA* duplex then separates. The strand more often loaded into the microRNA-induced silencing complex (miRISC) is referred to as mature miRNA or simply miRNA and the other, more often degraded, strand is miRNA*. Figure adapted from [21].

The mature miRNA acts as an adaptor. Through complementary base pairing with sequences in 3′ untranslated regions (3′UTRs) of particular protein coding transcripts in the cytoplasm, it guides the miRISC towards them [94, 95]. The binding of miRNA-miRISC to a protein coding transcript leads to degradation of the transcript and/or to inhibition of protein synthesis [95]. Therefore, the mechanistic effect of individual microRNAs is to downregulate their targets.

Regardless of which protein coding gene is targeted, the miRNA-miRISC binding to mRNA normally requires continuous base pairing between the 3′UTR and bases 2 to 7 of the mature miRNA (Figure 2.3). Additional match at nucleotide 8 is required if the mature miRNA sequence does not start with an A [96]. The 7 nucleotides at bases 2 to 8 are known as the seed region. Extensive complementarity in the rest of the sequence is unusual in animal miRNAs [96].

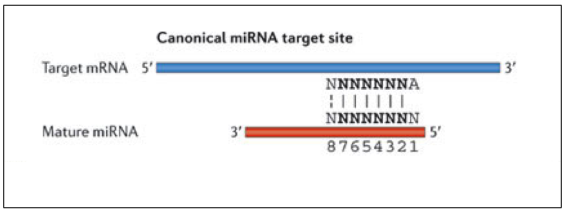


Fig. 2.3 **The canonical miRNA target site**. The mature miRNA guides miRISC to its target by complementary base pairing in the seed region. Figure adapted from [21].

2.2.2 MicroRNAs in vertebrate development and evolution - a hypothesis

On the whole, protein coding genes tend to evolve very slowly. We share almost all of our protein coding genes with the chimpanzee, and approximately 88% with rodents [97]. Many aspects of development are governed by transcription factors shared across the animal kingdom, from the fruit fly to human,and genes in evolutionarily distant organisms are remarkably similar, especially in functional terms, and often can be substituted for one another [98]. These and other observations reviewed by Caroll [98] provide a compelling argument that changes in gene regulation are likely to be important genetic drivers of adaptive differences between closely related animal species.

Regulatory changes affect development and can facilitate adaptive evolution by inducing differences in the timing or strength of gene expression. Many microRNAs play precisely these roles in development. Two first two microRNA genes that were identified, *lin-4* and *let-7*, were discovered by screening for genes that control developmental timing in *C. elegans* larvae [99, 100]. mir-10, a microRNA conserved from *C. elegans* to human, originates from genomic regions known to harbour the Hox genes, crucial regulators of animal development, and it is known to target mRNAs of several Hox genes to inhibit protein synthesis [101]. Finally, the extensiveness of microRNA involvement in development is exemplified for example by the haematopoiesis pathway, where “virtually every step seems to be fine-tuned by specific microRNAs” [102].

Many microRNAs also impact evolution by fine-tuning gene expression. There is growing evidence that this leads to added robustness in regulatory networks and increased phenotypic reproducibility in the face of environmental perturbations [103, 104, 105]. Such ‘canalisation’ of development increases heritability, making traits more responsive to natural selection [21]. For example, a comparative study of the localisation of ancient microRNAs in a sea anemone, marine worms, and a sea urchin revealed close connection between establishment of new tissues and new microRNAs in early bilaterian evolution [106].

Therefore,we hypothesise that a) differential miRNA regulation of protein coding genes may contribute to some of the phenotypic differences between cichlids; b) miRNAs, through increased canalisation, may contribute to the increased speed of evolution in some cichlid lineages.

2.3 MicroRNA annotation

2.3.1 Small RNA sequencing

To generate experimental evidence for microRNA gene annotation, I prepared small RNA sequencing libraries from late stage embryos of all five species whose genomes were assembled by the Broad Institute. All samples were obtained from the same strains of fish used for genome sequencing by the Broad Institute, and were staged according to developmental milestones set out in ref [107]. Stage ∼22 embryos (corresponding to length of approximately 7.2mm and age of 8 days post fertilisation in *O. niloticus*) were homogenised in TRIsure (Bioline) reagent and total RNA extracted using the manufacturer’s protocol. Using 5μg of total RNA as input, I used the Illumina TruSeq Small RNA Sample Preparation Kit to generate small RNA sequencing libraries, again following the manufacturer’s protocol. The libraries were sequenced on the Illumina MiSeq platform, yielding between 2.4 and 4.4 million 36bp single-end reads per sample.

Because the mature miRNA themselves are ∼22bp long, the ends of 36bp reads contained sequences from 3$'$ adaptors (specific sequences attached to the ends of the small RNA molecules to facilitate sequencing). The adaptor sequences were removed from reads using the `cutadapt v1.0` tool [108]. Figure 2.4 shows the length distribution of sequences after adaptor removal, and thus provides an indication about the proportion of reads in the dataset that are likely to correspond to sequences of mature miRNA molecules. The vast majority of reads came from molecules within the usual miRNA size range (20 to 24bp).

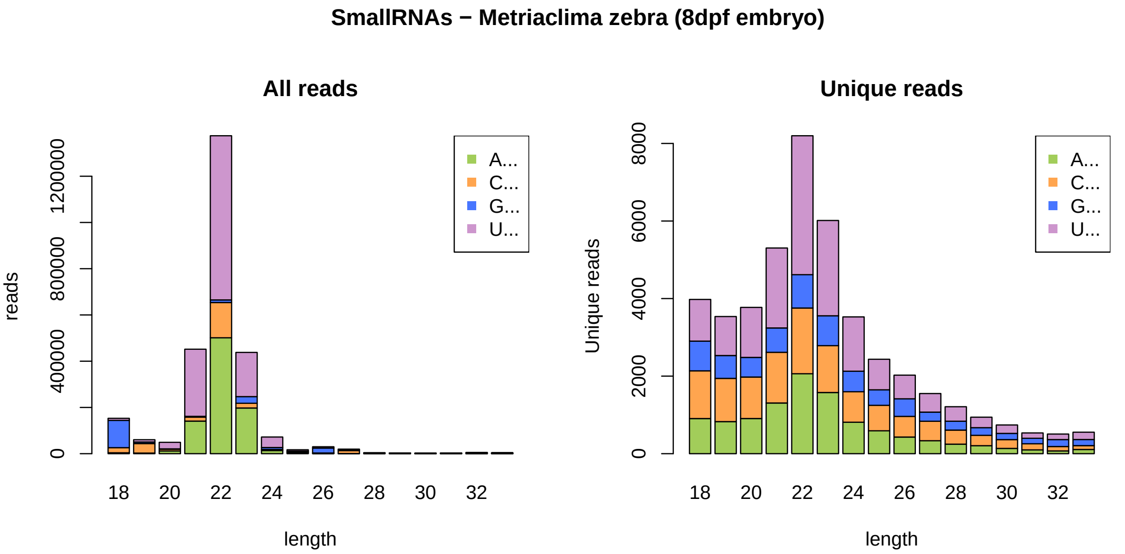


Fig. 2.4 **Length distribution of small RNAs sequenced from *M. zebra* embryos.** Colour denotes the first base of the read, as indicated in the legend. **Left:** All reads counted. The vast majority come from molecules within the usual miRNA size range. **Right:** Only unique reads counted. It is clear that the miRNA reads have low complexity - i.e. there are a lot of identical reads coming from the same miRNA.

2.3.2 Identification of miRNA loci

I used the `mirdeep2 v2.0.0.5` [109] package to detect signatures of microRNA genes from the sequencing data as follows. First, reads were ‘collapsed’, so that only one copy of identical sequences was kept, together with a record about how often the sequence is present in the data. Next, reads were mapped to their corresponding genome using the `bowtie v0.12.7` read aligner [110]. Reads mapping to multiple (more than five) genomic loci were discarded - these would be mostly repeat-derived sequences. Then looking at the remaining reads, the `RNAfold v2.0` program [111] was called to consider whether RNA derived from the sequence flanking any stacks of reads can plausibly fold into the hairpin structure, the signature of miRNAs. Finally, the localisation of reads within the predicted hairpin structure was evaluated. True miRNA reads should correspond predominantly to the mature 22nt sequence and should have a well aligned 5′ start site due to consistent processing by the Dicer enzyme. For a detailed description of the process, please see the `mirdeep2` publication by Friedländer *et al.* [109].

The `mirdeep2` pipeline also takes advantage of homology with known miRNAs from related organisms. Therefore, I prepared a file containing all experimentally validated teleost miRNAs present in `miRBase v.19` (`TELEOST_MATURE_miRNA.fa`). Then the mapper.pl and miRDeep2.pl scripts from the `mirdeep2` package were executed as follows:

```
mapper.pl READS_FILE -j -l 18 -m -p GENOME_ASSEMBLY -s reads_collapsed.fa
-t reads_collapsed_vs_genome.arf -v

miRDeep2.pl reads_collapsed.fa GENOME_ASSEMBLY reads_collapsed_vs_genome.arf
none TELEOST_MATURE_miRNA.fa none 2 > report_mirbase19.log
```

Using output from the above commands, I constructed a high confidence set of cichlid miRNA loci by selecting predicted loci that received `mirdeep2` score greater than or equal to 10 in any of the five species, except where fewer than 2% of the aligned reads were a perfect match to the predicted mature sequence (± one nucleotide at the 3′ end, mismatch at the 3′ end base allowed). For example, a miRNA locus with score < 10 in *M. zebra* would be included if its predicted mature sequence were identical to a miRNA with mirdeep2 score $>= 10$ in *O. niloticus* or another cichlid species.

In this way, I identified 1,344 microRNA (miRNA) loci (259 - 286 per species). The complete annotation has been submitted to `miRBase` [112], the central miRNA repository.

2.3.3 Evolution in cichlid miRNA repertoires

The ways miRNAs evolve and change their target repertoires are illustrated in Figure 2.5. I searched for sources of novelty in cichlid miRNA repertoires by comparison with known teleost miRNAs and I discovered a) 40 cases of de-novo miRNA emergence and nine cases of apparent miRNA loss (Figure 2.6a); b) four distinct mature miRNAs with direct mutation(s) in the seed sequence; c) at least 14 cases of arm switching; d) one case of seed shifting; e) 92 distinct miRNAs with mutation(s) outside the seed sequence. For detailed methods see section 2.4.1.

I shared my results with Jeffrey Streelman at the Georgia Institute of Technology, who used RNA *in situ* hybridisation experiments to explore the spatial expression

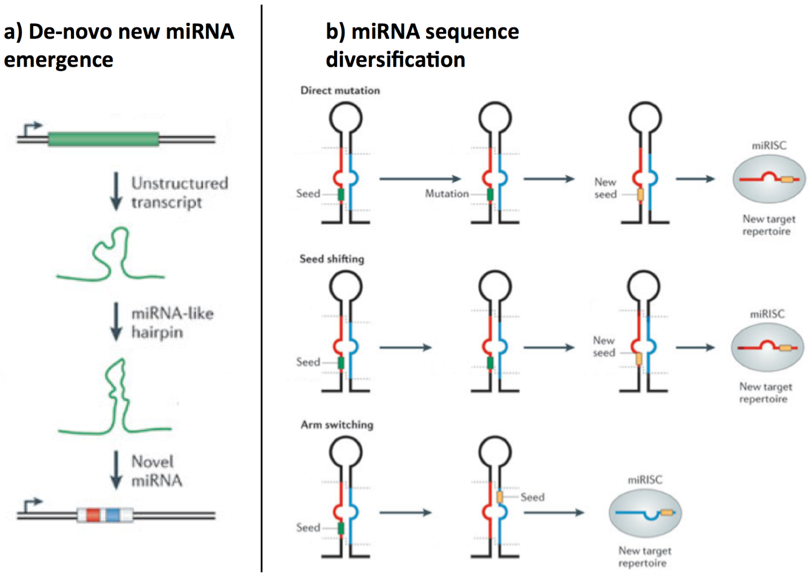


Fig. 2.5 **Evolution of miRNA novelty:** A miRNA with a new target repertoire can arise: **a)** de-novo from any transcribed sequence (e.g. intron); **b)** By diversification of an existing miRNA in the way of: 1) Direct mutation in the seed; 2) Change in hairpin processing which shifts the 5′ end of the miRNA (seed shifting); 3) Change in hairpin processing which leads to the previously degraded strand (blue) being utilised as a miRNA (arm switching); Figure from [21].

patterns in *M. zebra* for one case of arm switching (mze-miR-7132a-5p and mze-miR-7132a-3p). We found that spatial expression is clearly differentiated in the miRNA-miRNA* pair (Figure 2.6b), consistent with previous reports [113] and suggesting miRNA strand preferences may be controlled developmentally.

There is little comparable data on miRNA sequence novelty at similar evolutionary scales in other vertebrate lineages, but comparison with published data on *Drosophila* [114] suggests that the rate of de-novo miRNA emergence is not unusually elevated in cichlids. The cases of mutation in the seed sequence and of seed shifting are isolated incidents, too few to infer any trend. However, arm switching seems to be widespread (I have identified 15 cases but there are likely to be many more) and is likely the main mechanism by which cichlids generate miRNA sequence novelty. The evolution of miRNAs is intertwined with evolution of their targets, the protein coding genes they regulate. The detailed catalog of miRNA sequence novelty presented here provides a basis for exploration of the miRNA target space.

Novel cichlid miRNAs have complementary expression to predicted targets. I used the `RNAhybrid` [115] software package to predict genes that may be regulated by four de-novo cichlid-specific miRNAs: miR-10029, miR10032, miR-10044, miR-10049

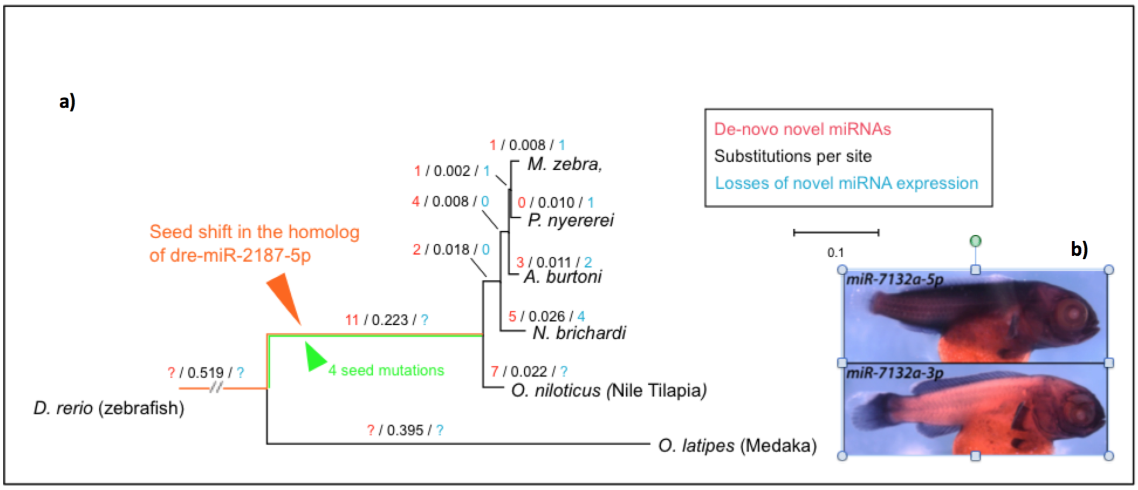


Fig. 2.6 **a)** The birth and death of de-novo novel microRNAs mapped on the phylogenetic tree of the five cichlid species. **b)** In situ hybridisation images showing the difference in spatial expression between mature miRNAs generated from 5p and 3p arms of the miR-7132a precursor in *M. zebra* (Jeffrey Streelman, Georgia Institute of Technology)

(detailed methods in section 2.4.2) and, again, shared the results with Jeffrey Streelman. His *in situ* hybridisation experiments revealed that, in *M. zebra*, the spatial expression of the four de novo miRNAs is confined to specific tissues (for example, fins, facial skeleton, brain), suggesting their expression is tightly regulated, and is strikingly complementary to genes predicted to contain target sites for these miRNAs (miR-10032 target *neurod2*, and miR-10029 target *bmpr1b*).

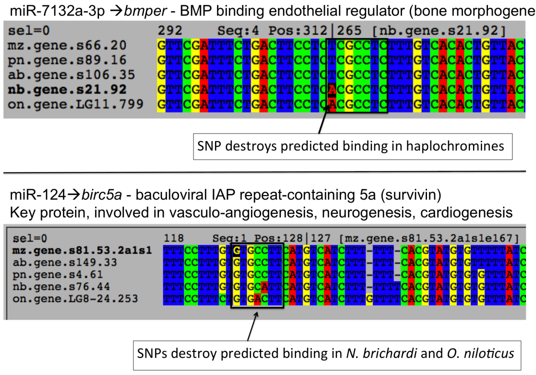
2.3.4 Evolution of miRNA targets

The target sequences bound by miRNAs tend to evolve much faster than miRNAs themselves [21]. To be able to observe sequence evolution at miRNA target sites across the East African cichlid radiation, I curated a set of 2359 genes with 1:1 homology in all five cichlid species and a unique 3′ untranslated region (3′UTR) sequence per gene per species. This was made possible by starting from the `V1` gene annotations generated at the Broad Institute as a part of the Cichlid genome project [93], which includes 3′UTRs and assignment of orthologs between the cichlid species. To ensure that any observed differences between species are due to substitutions or short insertions or deletions, I restricted the analyses to the portion of 3′UTRs present in all five species genome assemblies.

Fine scale evolution of particular miRNA targets

I focussed on the evolution of sites targeted by novel miRNAs, by miRNAs involved in arm switching in cichlids, and by 29 of the 38 miRNAs whose expression profiles in the central nervous system of zebrafish have been studied in detail by Kapsimali *et al.* [116]. I used the PACMIT software package [117] for target predictions (see section 2.4.3 for detailed methods), and focussed on identifying sites where sequence variation leads to gain or loss of predicted binding sites in *M. zebra*, *P. nyererei*, and *A. burtoni*, members of the most rapidly spectating cichlid tribe Haplochromini.

Using this strategy, I was able to identify several interesting cases where sequence evolution in 3′UTRs is predicted to create differences between the haplochromines and other East African cichlids in miRNA regulation of particular genes. For example, the mature miRNA produced from the 3p arm of the arm switching miR-7132a is predicted to target the key developmental gene *bmper* in *O. niloticus* and in *N. brichardi* but a substitution destroys the predicted binding site in haplochromines, possibly removing *bmper* from the control of this miRNA. Conversely, miR-124 is predicted to regulated the gene *birc5a* in haplochromines, but the binding site is lost in *O. niloticus* and in *N. brichardi*. This is interesting because in zebrafish *birc5a* promotes neuronal differentiation [118] and expression of miR-124 is associated with the transition of neural progenitors to differentiated neurons [116]. Therefore, the substitutions in *birc5a* 3′UTRs may result in differences in neural development and function between haplochromines and other East African cichlids.



miR-7132a-3p →*bmper* - BMP binding endothelial regulator (bone morphogenesis)

```
sel=0               292     Seq:4 Pos:312|265 [nb.gene.s21.92]
mz.gene.s66.20      GTTCGATTTCTGACTTCCTCTCGCCTCTTTGTCACACTGTTAC
pn.gene.s89.16      GTTCGATTTCTGACTTCCTCTCGCCTCTTTGTCACACTGTTAC
ab.gene.s106.35     GTTCGATTTCTGACTTCCTCTCGCCTCTTTGTCACACTGTTAC
nb.gene.s21.92      GTTCGATTTCTGACTTCCTCACGCCTCTTTGTCACACTGTTAC
on.gene.LG11.799    GTTCGATTTCTGACTTCCTCACGCCTCTTTGTCACACTGTTAC
```

SNP destroys predicted binding in haplochromines

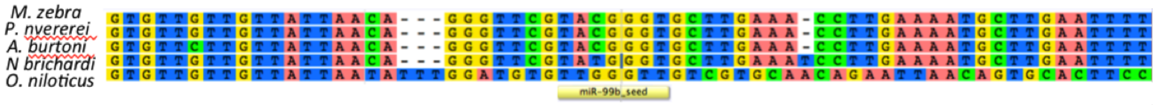
miR-124→*birc5a* - baculoviral IAP repeat-containing 5a (survivin)
Key protein, involved in vasculo-angiogenesis, neurogenesis, cardiogenesis

```
sel=0                  118     Seq:1 Pos:128|127 [mz.gene.s81.53.2a1s1e167]
mz.gene.s81.53.2a1s1   TTTCCTTTGTGTGCCTTCATGTCATCTTT-TTT-CACGTATGTGTTTTTATC
ab.gene.s149.33        TTTCCTTTGTGTGCCTTCATGTCATCTTT-TTT-CACGTATGTGTTTTTATC
pn.gene.s4.61          TTTCCTTTGTGTGCCTTCATGTCATCTTT-TTT-CACGTATGTGTGTTTATC
nb.gene.s76.44         TTTCCTTTGTGTGCATTCATGTCATCTTT-TTTTCACGTATGTGTGTTTATC
on.gene.LG8-24.253     TTTCCTTTCTGTGACTTCATGTCATCTTTGTTTTCACGTATGTGTGTTTATC
```

SNPs destroy predicted binding in *N. brichardi* and *O. niloticus*

Fig. 2.7 **Sequence evolution at specific miRNA target sites**. Multiple alignments of 3′UTRs of: mz-*M. zebra*, pn-*P. nyererei*, ab-*A. burtoni*, nb-*N. brichardi*, and on-*O. niloticus*. Predicted target sites (sequences complementary to miRNA seeds) are denoted by black rectangles and substitutions (SNPs) predicted to alter binding are highlighted.

The way a miRNA targets a gene for regulation is still imperfectly understood and miRNA target prediction algorithms all suffer from large numbers of false positives (see section 2.4.3). However, it is possible to experimentally illustrate that a miRNA is interacting with a transcript, for example by showing that it can reduce the level of protein expression in an *in vitro* system using a luciferase assay [119]. My target predictions provided a basis for such an experiment. The experiment was conducted by Joe Hanly, a research project student at the Miska Laboratory at the Gurdon Institute in spring 2013. He used the psiCHECK-2 vector[1] to assay the ability of miR-99b to downregulate the expression of the gene *tmem141*. My computational predictions suggested that the gene should be targeted in *M. zebra* and other haplochromines, but that the binding site is not present in *O. niloticus* and *N. brichardi* (Figure 2.8).



M. zebra	GTGTTGTTGTTATTAACA---GGGTTCGTACGGGTGCTTGAAA-CCTTGAAAATGCTTGAATTTT
P. nyererei	GTGTTGTTGTTATTAACA---GGGTTCGTACGGGTGCTTGAAA-CCTTGAAAATGCTTGAATTTT
A. burtoni	GTGTTCTTGTTATTAACA---GGGTTCGTACGGGTGCTTGAAA-CCTTGAAAATGCTTGAATTTT
N brichardi	GTGTTGTTGTTATTAACA---GGGTTCGTATGGGTGCTTGAAATCCTTGAAAATGCTTGAATTTT
O. niloticus	GTGTTGTTGTTATTAATATTTGGATGTGTTGGGTTGTCGTGCAACAGAATTAACAGTGCACTTCC

Fig. 2.8 **Sequence evolution in *tmem141* 3′UTRs** The predicted target site (sequence complementary to the seed sequence of miR-99b) is delineated by the yellow rectangle below the sequences. (Figure by Joe Hanly, research project student at Miska Lab in spring 2013)

3′UTRs of *tmem141* from *M. zebra*, *O. niloticus*, and *N. brichardi* were cloned into the psiCHECK-2 luciferase assay vector and transfected into human cell culture. Upon introduction of miR-99b, the activity of luciferase protein with the *M. zebra* 3′UTR was significantly reduced by 20% +/- 7.6% compared to control (Figure 2.9), while luciferase constructs with *O. niloticus* and *N. brichardi* 3′UTRs were not significantly downregulated (Figure 2.9). The function and expression profile of *tmem141* in fish is not known, and the gene was selected for this experiment by Joe Hanly simply for ease of 3′UTR cloning. Nevertheless, these results provide experimental evidence complementary to and in agreement with my computational predictions and constitute a first attempt at experimental validation of a miRNA target in cichlid fishes.

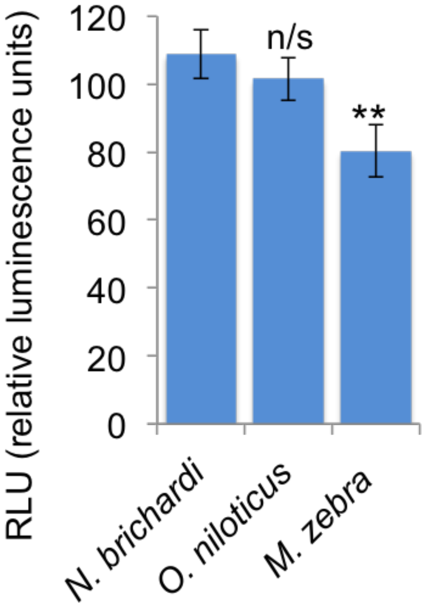


Fig. 2.9 **miR-99b downregulates luciferase with *M. zebra* *tmem141* 3′UTR**.

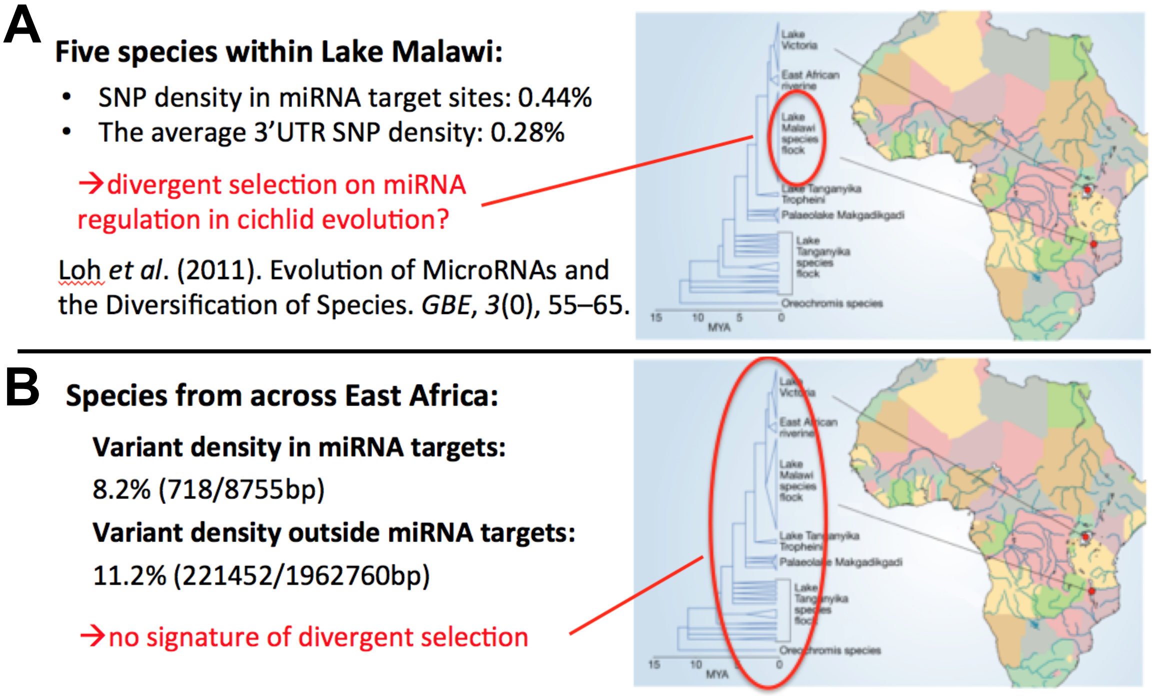
---

[1] https://www.promega.co.uk/resources/pubhub/cellnotes/microrna-biosensors-application-for-the-psicheck2-vector/

Increased conservation of target sequences across the East African radiation

In 2011, Loh, Yi, and Streelman conducted a study on low coverage genomic data from five Lake Malawi species and concluded that there is evidence of "accelerated divergence of miRNA target sites in cichlids, suggesting that the selective divergence of miRNA regulation has a role in the diversification of cichlid species." [120]. Specifically, the authors found that the density of single nucleotide variants in predicted miRNA targets was 0.44%, much higher than the average 3′UTR density of 0.28% (Figure 2.10A). While this is a very interesting finding, I note that the study relied on limited data: only 11.6Mb of multiple sequence alignment (only 1.6% of the assembled *M. zebra* genome), only computational predictions of miRNAs and protein coding genes, and an arbitrary decision that all 3′UTRs extend exactly 500bp from the end of the protein coding sequence.

With the whole genome assemblies from the Cichlid Genome Consortium and benefitting from experimental miRNA and 3′UTR annotations, I was able to confidently test whether accelerated divergence of miRNA target sites can also be observed across the broader East African radiation. I used all *M. zebra* miRNAs and all annotated 3′UTRs in my curated a set of 2359 genes to predict miRNA target sites with the PACMIT software package [117] (see section 2.4.3 for detailed methods). Then I generated multiple alignments of the 3′UTR sequences with `clustax v2.0` and found that the density of single nucleotide variants within predicted miRNA targets is 8.2% (718/8,755bp) while the density over the rest of 3′UTR sequences is 11.2% (221,452/1,962,760bp) (Figure 2.10B). Therefore, these findings contrasts with that of Loh, Yi, and Streelman [120]. I conclude that there is evidence of elevated purifying selection on miRNA targets over the evolutionary timescales that separate *O. nilotics*, *N. brichardi*, *A. burtoni*, *P. nyererei*, and *M. zebra*, consistent with functional constraint as in other functional sequence such as enhancers or protein coding sequence.



A **Five species within Lake Malawi:**

- SNP density in miRNA target sites: 0.44%
- The average 3’UTR SNP density: 0.28%

→divergent selection on miRNA regulation in cichlid evolution?

Loh *et al*. (2011). Evolution of MicroRNAs and the Diversification of Species. *GBE*, *3*(0), 55–65.

B **Species from across East Africa:**

Variant density in miRNA targets:
8.2% (718/8755bp)
Variant density outside miRNA targets:
11.2% (221452/1962760bp)

→no signature of divergent selection

Fig. 2.10 **Purifying selection on miRNA target sites**. Contrasting results were obtained by two studies: **(A)** Loh *et al.* found accelerated target site divergence in five Lake Malawi species. **(B)** I found evidence of elevated purifying selection on miRNA targets when comparing the five species sequenced by the Cichlid genome consortium. The maps and phylogenetic trees used in this figure are from [121].

2.4 Detailed methods

2.4.1 Searching for novelty in cichlid miRNA repertoires

I used programs from the `fasta v36.3.4` software package to align the high confidence set of cichlid mature and precursor (hairpin) miRNA sequences to all experimentally validated teleost miRNAs present in `miRBase v.19`. Specifically, I used the `ssearch` (local Smith-Waterman) algorithm for hairpin→hairpin and mature→mature alignments and the `glsearch` (global-local Needleman-Wunsch) algorithm for mature→hairpin alignment as follows:

```
> ssearch36 -E 0.005 -C 25 -m 9 -3 CICHLID_HAIRPIN.fa TELEOST_HAIRPIN.fa >
HAIRPIN_TO_HAIRPIN.ssalign

> glsearch36 -E 0.1 -C 25 -m 9 -3 CICHLID_MATURE.fa TELEOST_HAIRPIN.fa >
MATURE_TO_HAIRPIN.glalign

> ssearch36 -E 0.1 -C 25 -m 9 -3 CICHLID_MATURE.fa TELEOST_MATURE.fa >
MATURE_TO_MATURE.ssalign
```

Based on analysing the alignments, I used a custom script to automatically assign all cichlid miRNAs into five categories based on their conservation/novelty. The name of the file in GREEN colour indicates presence of at least one significant alignment for a cichlid miRNA in miRBase v.19; RED colour indicates absence of any significant alignment.

1) HAIRPIN_TO_HAIRPIN+MATURE_TO_HAIRPIN+MATURE_TO_MATURE

These are homologs of known teleost miRNAs, possibly with single base polymorphisms:

a) Mature sequence identical to the best match (ignoring ± 2bp length difference at $3'$ end)
b) Mature sequence different from the best match but seed sequence (bases 2-8) identical
c) Mature sequence different from the best match with a difference in the seed
d) $5'$ isomiRs, leading to seed shifting

2) HAIRPIN_TO_HAIRPIN+MATURE_TO_HAIRPIN+MATURE_TO_MATURE

Alternative processing of a known hairpin - arm switching

3) HAIRPIN_TO_HAIRPIN+MATURE_TO_HAIRPIN

There is a homologous hairpin but the mature miRNA has changed substantially. This could be a sign of divergent evolution of the mature miRNA sequence.

4) HAIRPIN_TO_HAIRPIN+MATURE_TO_MATURE

There is no significant alignment for the hairpin but the mature sequence is highly similar to known teleost miRNA. This could be a sign of convergent evolution of mature miRNA sequence.

5) HAIRPIN_TO_HAIRPIN+MATURE_TO_MATURE

In this category are novel miRNAs likely to have arisen de-novo since the divergence between cichlid and medaka ancestors.

2.4.2 Target prediction by `RNAhybrid`

miRNA target prediction for the Cichlid genome consortium publication was done using `RNAhybrid` [115]. For each miRNA, the calculation of p-values was calibrated using the tool `RNAcalibrate` before using the `RNAhybrid` tool for the calculation of minimum free energy hybridisation between the miRNA and all annotated 3′UTRs in all five species. For both `RNAcalibrate` and `RNAhybrid` I forced perfect matches between bases 2 and 7 of the miRNA (the seed) using the `-f 2,7` option.

2.4.3 Target prediction with `PACMIT`

The false discovery rate for computational target prediction is often quoted to be approximately 50% [122]. It is difficult to to choose the best among the different methods because of a lack of studies comparing the available algorithms using a set of experimentally validated miRNA targets. Marín and Vanícek conducted one of only a few such studies [117], comparing several prediction methods on a set of experimentally validated miRNA targets in human (Figure 2.11). It is on the basis of this study that I chose the 'Prediction of Accessible MicroRNA Targets (`PACMIT`)' algorithm for miRNA target prediction analyses outside the Cichlid Genome Consortium.

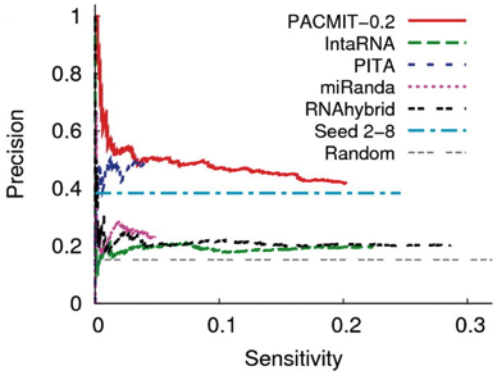


Fig. 2.11 **Precision versus sensitivity of a selection of miRNA target prediction algorithms**. Figure from [117].

`PACMIT` uses predictions of secondary structure of the $3'$UTR sequence to determine whether any locus with sequence complementarity to the miRNA seed would be accessible for binding. I used the ViennaRNA Package [111] to predict the secondary structures, excluding UTRs shorter than 22bp and also UTRs with undetermined bases (`N` characters). With the accessibility information available, I executed PACMIT to always require perfect pairing between bases 2-8 of the miRNA and the $3'$UTR and accessibility of the bases at miRNA positions 2-5, as suggested by the authors in personal communication and in [123] as follows:

```
paccmit.pl -utrs UTRs.fa -mirs miRNAs.fa -prop 3 -nuclend 5 -nmer 4 -pcutoff 0.2
-output out.txt
```